SeaLLMs - Large Language Models for Southeast Asia

Year: 2023 | *Citations: 112* | *Authors: Xuan-Phi Nguyen, Wenxuan Zhang, Xin Li*

Abstract

Despite the remarkable achievements of large language models (LLMs) in various tasks, there remains a linguistic bias that favors high-resource languages, such as English, often at the expense of low-resource and regional languages. To address this imbalance, we introduce SeaLLMs, an innovative series of language models that specifically focuses on Southeast Asian (SEA) languages. SeaLLMs are built upon the Llama-2 model and further advanced through continued pre-training with an extended vocabulary, specialized instruction and alignment tuning to better capture the intricacies of regional languages. This allows them to respect and reflect local cultural norms, customs, stylistic preferences, and legal considerations. Our comprehensive evaluation demonstrates that SeaLLM-13b models exhibit superior performance across a wide spectrum of linguistic tasks and assistant-style instruction-following capabilities relative to comparable open-source models. Moreover, they outperform ChatGPT-3.5 in non-Latin languages, such as Thai, Khmer, Lao, and Burmese, by large margins while remaining lightweight and cost-effective to operate.